Trait-unconsciousness, State-unconsciousness, Preconsciousness, and Social Miscalibration in the Context of Implicit Evaluations [1]

Adam Hahn
University of Cologne &
University of Bath

Alexandra Goedderz
University of Cologne

Implicit evaluations are often assumed to reflect "unconscious attitudes". We review data from our lab to conclude that the truth of this statement depends on how one defines "unconscious". A trait definition of unconscious according to which implicit evaluations reflect cognitions that are introspectively inaccessible at all times appears to be inaccurate. However, when unconscious is defined as a state in which cognitions can be in at specific times, some data suggest that the cognitions reflected on implicit evaluations may sometimes unfold without direct awareness in that people seem to rarely pay attention to them. Additionally, people appear to be miscalibrated in their reports in that they construe even conscious biases in self-serving ways. This analysis suggests that implicit evaluations do not reflect unconscious cognitions per se, but awareness-independent cognitions that are often preconscious and miscalibrated. Discussion centers on the meaning of this analysis for theory and application.

Keywords: Implicit bias, awareness, consciousness, calibration, attitudes, evaluations

In trying to distinguish implicit from explicit measures, one of the most prominent features has been the idea that implicit measures reflect "unconscious" cognitions in contrast to explicit measures, which reflect "conscious" cognitions (e.g., Greenwald & Banaji, 1995; Lai et al., 2013; McConnell et al., 2011; Nosek et al., 2002)[2]. Among other things, this idea is often based on the observation that implicit and explicit measures of the same targets generally show low correlations (Hofmann, Gawronski et al., 2005; Nosek, 2005; Nosek & Hansen, 2008); and it has been met with enthusiasm by the public. Different writers have declared the existence of "unconscious racisms" (Quillian, 2008) and "unconscious prejudice" (Powell, 2016) on the basis of research with implicit measures. The idea was challenged by data showing that people can predict the patterns of their implicit evaluations (Hahn et al., 2014; Hahn & Gawronski, 2019; Hahn & Goedderz, in prep. a, in prep. b). These data are in line with dual-process models ascertaining that the cognitions reflected on implicit measures are often consciously rejected for explicit reports, for example, because people are "unwilling" to admit to

[1] This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors permission. The final article will be available upon publication. Cite as: Hahn, A., & Goedderz, A. (in press). Trait-unconsciousness, State-unconsciousness, Preconsciousness, and Social Miscalibration in the Context of Implicit Evaluations. *Social Cognition.*

Adam Hahn at the University of Bath, the United Kingdon and Alexandra Goedderz, University of Cologne, Germany. Correspondence should be addressed to Adam.Hahn@uni-koeln.de

[2] We use the terms "implicit" and "explicit" on the level of measurement outcomes (i.e., "measures"). According to this definition, an "implicit evaluation" is an evaluation inferred from indirect computerized measurement instruments, whereas an "explicit evaluation" is an evaluation stated in self-report measures (Hahn & Gawonski, (2018)). These definitions make no inferences about the underlying cognitions that are reflected on each type of measurement outcome. Instead, we use the current paper to weigh the different implications the data have for understanding those underlying cognitions.

them (ProjectImplicit.net, 2020). Such models present these cognitions as conscious, but rejected (Fazio, 2007; Gawronski & Bodenhausen, 2006).

At the same time, the idea that implicit measures reflect fully conscious cognitions appears to stand at odds with other observations. For instance, people are surprised and defensive at IAT feedback (Gawronski, 2019; Goedderz & Hahn, in prep.; Howell et al., 2013; Howell et al., 2017; Schlachter & Rolf, 2017), and most people seem to think that they are less biased than others, a statistically impossible perception (Hahn et al., 2014; Howell & Ratliff, 2017). Such observations suggest that there are things people do not know about the cognitions reflected in implicit evaluations, and this may justify calling them "unconscious".

In the present paper, we will review evidence for both arguments as they pertain specifically to implicit evaluations. In doing so, we argue, first, that whether or not the cognitions reflected on implicit evaluations should be called "unconscious" depends on how one defines this term (Norman, 2010). Whereas defining a whole class of cognitions (e.g., all cognitions reflected on implicit measures) as unconscious implies a *trait* definition of unconsciousness, most researchers and lay people appear to think of unconsciousness as a *state* of a specific cognition that can change. A second problem is methodology. Which analyses would demonstrate whether a sample of study participants is aware or unaware of the cognitions showing on implicit measures? We argue that introducing a new construct – "social calibration" – can help explain contradictory findings in this domain. Distinguishing between social calibration and awareness can demonstrate (1) how people who think they are less biased than others may harbor a lack of social awareness, but still be introspectively aware of their biases; and (2) why low between-subjects correlations between reports on cognitions and implicit measures might be misleading when interpreted as evidence for unconsciousness.

Below, we start by explaining each of the constructs and definitions in detail, including what empirical observations would support calling the cognitions reflected on implicit measures unconscious under each definition or construct. We will then present some ongoing research from our lab in an attempt to answer whether implicit evaluations reflect trait-unconscious, state-unconscious, and/or frequently miscalibrated cognitions. We conclude by discussing the meaning of these findings to dual-process models and usage of implicit measures in applied contexts.

Definitions

We use the terms conscious and unconscious as attributes for cognitions, and the terms aware and unaware to describe the hosts of those cognitions. Hence, a cognition of which a person is unaware is unconscious, whereas a cognition of which a person is aware is conscious.[3] Gawronski et al. (2006) distinguish between three aspects of an automatic cognition of which a person can be aware: Its source (i.e., how did the cognition develop?), its content (the cognition itself), and the impact it has on behavior. Our analysis is focused on content. In other words, we ask whether people are aware of the cognitions reflected on implicit measures.

[3] Some researchers use the term "awareness" as an attribute for both people and cognitions (e.g., Hütter et al. (2012); Hütter and Sweldens (2013)). To address the frequent usage of the term "unconscious attitudes," we use the terms "conscious" and "unconscious" to refer to cognitions in this article.

Based on previous work on consciousness, we further assume that whether or not a cognition reaches conscious awareness depends on (1a) the strength of the signal it produces, and (1b) the degree of attention that is directed towards this signal (Dehaene & Naccache, 2001; Hahn & Goedderz, in prep. b; Hofmann & Wilson, 2010).

Unconscious Definition 1: Inaccessible to Introspection (Trait-unconscious)

This definition implies that an unconscious cognition is *impossible* to know for its host unless this host is informed of its existence from another source. For instance, it is impossible to observe how the reflections of light on the retinae of our eyes initially project images that are upside-down, and only later transformed into three-dimensional, upright images of our environment. We may learn about this process from books and scientists, but it is impossible to observe it unfold directly via introspection. The process is trait-unconscious. Regarding the two processes introduced in the beginning, a trait-unconscious cognition refers primarily to Process 1a: It produces no detectable signal. Regardless how much attention is paid, there is nothing a person could observe that would allow them to notice the cognition.[4]

A trait definition of "unconscious" in the term "unconscious attitudes" implies that people would be *unable* to report on the contents of the cognitions reflected in implicit evaluations without completion of a test and feedback about their performance. It is prominent in the first half of the popular definition of implicit evaluations as "attitudes people are unable or unwilling to report" (ProjectImplicit.net, 2020), and many authors have explicitly made claims about inaccessibility (e.g., Kassin et al., 2011; Kihlstrom, 2004; Nosek, 2005, 2007). It stands at odds with dual-process models that assert that the cognitions reflected on implicit measures are consciously rejected when people decide what to report on explicit measures (Fazio, 2007; Gawronski & Bodenhausen, 2006). These models define the cognitions reflected in implicit evaluations as trait-conscious. Empirically, the trait-unconsciousness assumption is challenged by findings that people can predict the patterns of their implicit evaluations (Hahn et al., 2014; Hahn & Gawronski, 2019; Hahn & Goedderz, in prep. a). We discuss this research and whether implicit measures reflect trait-unconscious cognitions in Question 1.

Unconscious Definition 2: Not Residing in Awareness at a Specific Point in Time (State-Unconscious or Preconscious)

You may have been in a situation in which you noticed for the first time that you disliked the smell of a specific food item, and you felt that you had been "unaware" of this evaluation beforehand. We would argue that you were using a state definition of awareness in this context; your evaluation was "state-unconscious" until you paid attention to it. It may be described as a "state" definition, because here unconsciousness is a state of the cognition that defined it only at a specific time (before you thought about it), and this state changed as soon as you directed your attention to it. State-unconsciousness can happen at both Stages 1a and 1b. Specifically, moderating conditions may influence the strength of a signal a cognition produces (Process 1a), such that it remains state-unconscious until those conditions are

[4] Applied to Process 1b, a class of cognitions would be trait-unconscious if no person ever paid attention to the signals the cognitions produce, even though such signals exist. We consider this option hypothetical and will not discuss it in the present article.

present. For instance, you may be truly unaware that you dislike a specific smell until you are confronted with a concrete instance of it. At the same time, you may remain unaware of your reaction simply from not paying attention even when the smell is present (Process 1b). When cognitions are known to be accessible when attention is paid, but remain unconscious as long as a person does not pay attention, they may also be called "preconscious" in the pre-attention state (Dehaene et al., 2006).

We have divided our review of the evidence of state-unconsciousness and preconsciousness into two questions mapping onto the two processes. In Question 2a, we ask whether independent moderators can facilitate or inhibit the strength of the signal the cognitions reflected on implicit measures produce. If, holding attention constant, certain factors facilitate accurate prediction of IAT scores, then this would indicate that the cognitions reflected on implicit evaluations can be state-unconscious when those conditions are absent. Question 2b asks whether people remain unaware of their biases until they are asked to pay attention to them. A negation of this question (people do *not* pay attention until asked) would imply that the cognitions reflected in implicit measures are preconscious for many people a lot of the time.

Note that these points would not justify calling the cognitions reflected on implicit measures unconscious per se. Instead, they would suggest that the cognitions reflected on implicit evaluations are independent of awareness, such that they can variably unfold in a conscious, preconscious, or unconscious state.

Social Calibration

Social calibration describes how consistently different people use the same labels to describe the same cognitions. To understand the concept of social calibration, imagine you are asked to predict how you would score on a racial bias IAT. Now imagine further you had to rate how strong exactly your bias is. Are you going to show a "slightly more positive reaction" towards one group than another, or a "strongly more positive reaction"? To be able to answer this question, you would have to know (1) how other people perceive "slight" or "strong" biases, and (2) be able and willing to apply this standard to your own cognitions (Hahn & Goedderz, in prep. b). The difference between knowing you are biased and knowing what to call your bias in comparison with others is what we refer to when we distinguish social calibration from introspective awareness (Hahn & Goedderz, in prep. b). Introspective awareness describes whether or not a person is aware of a cognition, whereas social calibration describes whether they are aware of where this cognition lies in comparison to a reference sample.

Social calibration matters in discussions around awareness for at least two reasons. First, it can explain why people may have unrealistic beliefs that they are less biased than others, even though they are aware of being biased per se. If a person believes that they are less biased than other people perceive them to be, this can be indicative of a lack of social awareness – similar to overlooking one's impoliteness or lack of kindness as it is perceived by others. As we will show, however, it is different from failing to notice a cognitive process introspectively. Second, awareness and calibration are systematically confounded in between-subjects correlations between reports of a cognition (e.g., prediction of an IAT score) and criteria for the same cognition (e.g., actual IAT scores), an issue to which we turn next. We discuss empirical evidence on whether people are socially calibrated in their attitude reports in Question 3.

Distinguishing introspective awareness from social calibration methodologically. Most research on awareness infers its presence or absence from between-subjects correlations (e.g., Hofmann et al., 2009, 2005; Howell et al., 2013). For instance, Howell et al. (2013) found a non-significant correlation of .12 between IAT score predictions and actual scores on a Black-White IAT. Most participants expected to show lower biases (e.g., a "slight preference") than their feedback indicated (e.g., a "strong preference"). Importantly, for between-subjects correlations between predicted and actual IAT scores to be high, participants have to agree on how bias scores are labeled. The most biased person in a sample would have to agree to use the end of a scale, and the person with the weakest biases would have to agree that their biases should be labeled towards the center of the scale, close to "no bias". They would have to be socially calibrated.[5] While it is true that the participants in Howell et al.'s (2013) study may have been unaware of their biases, a between-subjects analysis simply doesn't allow researchers to distinguish between awareness and calibration.

To assess only introspective awareness, analyses have to be conducted within-subjects across targets. For within-subjects correlations to be high, participants in a sample would have to be able to say whether or not their individual biases towards, e.g., Latinos is stronger than their biases towards Asians, and different again from their biases towards other social groups. The patterns they describe should map onto the patterns of scores they show, independent of whether they believe all of those biases are milder or stronger than the biases of others.

In summation, within-subjects correlations per participants across targets can demonstrate introspective awareness, whereas between-subjects correlations per target across participants speak to how well-calibrated participants are in their reports on responses to individual target groups.

Questions Pertaining to Awareness and Calibration in Implicit Measures

Question 1: Are the Cognitions Reflected in Implicit Evaluations Trait-unconscious (i.e., Inaccessible to Conscious Introspection)?

Before Hahn et al. (2014), most research on awareness had looked at between-subjects correlations between implicit and traditional explicit evaluations of the same targets (e.g., Hofmann et al., 2009, 2005; Nosek, 2005, 2007). A traditional explicit attitude question can range from "how warmly do you feel towards group X" to the Modern Racism Scale (McConahay, 1986, used to dissociate implicit and explicit evaluations, e.g., by Nier, 2005). Such between-subjects implicit-explicit correlations tend to be low. (e.g., r = .2 in a meta-analysis by Hofmann, Gawronski et al., 2005), and this was interpreted such that awareness of the cognitions reflected on implicit measures must also be low (Nosek, 2005; Nosek & Hansen, 2008). Importantly, however, a variety of dual-process models explain relations between implicit and explicit evaluations differently (Devine, 1989; Gawronski & Bodenhausen, 2006; Rydell &

[5] This is especially true for homogenous samples, such as White undergraduates from the same university, who can be expected to show similar levels of bias against certain groups (Payne et al. , 2017). More diverse samples could show higher between-subjects correlations simply from distinguishing people who are biased in favor of one group from people who are biased in favor the other group, even if the participants disagree on how to label those biases.

McConnell, 2006). For instance, according to both the Associative-Propositional Evaluations Model (APE, Gawronski & Bodenhausen, 2006) and the Motivation and Opportunity as Determinants Model (MODE, Fazio, 2007), implicit evaluations reflect associations between an attitude target and valence. The models further assume that these associations are often rejected or overridden when people are asked for an explicit evaluation. According to the MODE model, explicit evaluations will differ when a person is motivated and has the opportunity to override the initial response implied by an association. According to the APE model, implicit evaluations are rejected when the person considers other information that has opposing evaluative implications more valid. For instance, a White American who is asked to evaluate African Americans might think of their Black friends and colleagues, other admired Black exemplars, and/or their egalitarian values. Even if the person harbors negative associations with African Americans as a category, they may reject these associations for their report in favor of this other, more positive, information. According to these models, implicit-explicit correlations can show whether people consider the cognitions reflected on implicit measures valid for judgment, not whether they are aware of them.

Integrating these thoughts, Hahn et al. (2014) asked participants to predict how they would score on five upcoming IATs. Table 1 shows results averaged from two national contexts. Prediction accuracy was calculated within-subjects across five predictions and five IATs. It was always high, even as explicit measures showed lower correlations with the same IAT scores. The authors additionally looked at how much of their own pattern each person could predict over and above a random other participant (and hence beyond a normative pattern). Analyses supported the notion that people have unique insight into the patterns of their own scores. Relations between implicit and explicit evaluations of the same targets could furthermore be explained entirely by IAT score predictions (see Table 1). These studies support the APE and MODE models' notions that people can be aware of the cognitions reflected on implicit measures, but may often reject them as bases for explicit reports. In sum, according to our data the answer to Question 1 is no – the cognitions reflected in implicit evaluations are not inaccessible, they are not trait-unconscious.

Question 2a: Do the Cognitions Reflected on Implicit Measures Remain (State-) Unconscious Until Specific Conditions Are Met?

In an ongoing project, Hahn and Goedderz (in prep. a) tested whether seeing concrete stimuli or knowledge of measurement had the bigger influence on prediction accuracy. In these studies, attention to spontaneous reactions (Process 1b) was held constant (participants were always asked to report on their spontaneous reactions to social groups), while two other factors in the prediction procedure were manipulated. Results favor concrete stimuli over knowledge of measurement as a determining factor. Participants who predicted their reactions to a hypothetical test, or simply reported their spontaneous reactions without announcement of a test, did not differ significantly from participants who were told that their reactions would be tested later. However, participants were less accurate without pictures. A follow-up study further showed that predictions were still more accurate when participants saw different pictures of the groups than those that would be used on the IATs. This latter finding suggests that the effect cannot be explained as a stimulus effect alone. That is, participants should always be more accurate when using the exact same stimuli for the

predictions due to stimulus-specific reactions, and the data supported this notion. However, the fact that participants tended to be better with other concrete pictures as well means that it must be something about the concreteness of the stimuli, and not just the specific stimuli themselves, that is responsible for the effect. One interpretation of these findings is that people may continue to recruit different propositional information (e.g., specific exemplars, values) when they predict a reaction in the abstract than when they predict it while looking at concrete stimuli. This interpretation is supported by the observation that participants' predictions were more strongly related to their explicit evaluations without pictures than with pictures (Hahn & Goedderz, in prep. a).

The idea that people need concrete stimuli to discover their biases is also supported by findings that people were less surprised at IAT scores when they first observed their reactions towards pictures, while making an abstract IAT score prediction without pictures did not reduce surprise (Goedderz & Hahn, in prep.). More research is needed to determine whether, and if so, which, moderating factors facilitate IAT score prediction accuracy when attention to spontaneous reactions is held constant. As it stands, the abstractness interpretation of the findings presented here suggests that the biases reflected on implicit measures may often be state-unconscious until people are confronted with concrete stimuli that trigger affective responses.

Question 2b: Do the Cognitions Reflected on Implicit Measures Remain Preconscious until People are Specifically Encouraged to Pay Attention to Them?

To answer this question, Hahn and Gawronski (2019) looked at changes on explicit measures after IAT score prediction. If people change their explicit evaluations, then this would suggest that they discover new information they had not considered previously. However, if people are always aware of their biases and consciously reject them as bases for explicit judgments, then those explicit judgments should not change in reaction to predicting IAT scores. Confirming the first hypothesis, results showed that the simple act of observing one's reactions and then predicting how one would score on IATs led to changes on explicit evaluations. Completing IATs – explained as a measure of implicit racial attitudes – had no effect by itself. This latter effect disconfirms a social-desirability explanation according to which participants simply gave up their covers and became "more honest" once they knew their biases would be assessed. If this were the case, IAT completion should have had the same effect as IAT score prediction. Later studies in the same article found that, even without reference to the existence of a test, encouragement to observe their reactions led participants to describe themselves as more biased.

Additional evidence comes from Goedderz and Hahn (in prep.). First, these studies confirmed the so-far anecdotal observation that people are more surprised when their IAT scores indicate bias than when they do not. However, here too, a simple encouragement to observe one's reactions to pictures of Black and White people reduced surprise at bias feedback on a Black-White IAT.

These studies suggest that people do not always consider the cognitions reflected in implicit evaluations and deliberately reject them. Instead, people seem to discover new information when they are asked to observe their own affective reactions towards concrete stimuli. IATs include concrete stimuli but no encouragement to pay attention to one's reactions. Accordingly, the fact that IAT completion alone leads people to react to feedback with surprise (Goedderz & Hahn, in prep.), defensive responses (Howell & Ratliff, 2017), as well as to

unchanged explicit evaluations and unchanged acknowledgment of bias compared to control (Hahn & Gawronski, 2019), supports the notion that the biases reflected on implicit measures are often preconscious until specific encouragement to pay attention.

Question 3: Are People Socially Calibrated When Reporting on the Cognitions Reflected on Implicit Evaluations?

To assess introspective awareness independent of social calibration, Hahn et al. (2014; see also Hahn & Gawronski, 2019; Hahn & Goedderz, in prep. b) calculated prediction accuracy within-subjects. Relevant for the current question, the same data can be used to look at how well-calibrated participants are by calculating between-subjects correlations between predictions and IAT scores for each attitude target pair. Between-subjects correlations between predictions and IAT scores for social-group biases were on average .31 across four studies on US-American undergraduates in Hahn et al. (2014), and .21 in Hahn and Goedderz (in prep. b) on German student participants. Additionally, Hahn et al. (2014) found that most participants thought that other participants in the same study would show more bias than they would show themselves, a statistically impossible result. Hence, participants in these studies seemed to be socially miscalibrated, even though they otherwise showed awareness.

One may argue that all calibration is arbitrary and there are no consensual cultural standards from which a person could learn to be better calibrated. Research on anchoring and adjustment has shown that people's judgments are often anchored upon arbitrary, situationally accessible information (Tversky & Kahneman, 1974). For instance, numerical judgments, even in domains as important as legal decisions, can be influenced by anchors as arbitrary as a role of a die (Englich et al., 2006). Contradictory to this perspective is the observation that humans talk about their preferences (for food, music, clothes) all the time. Both researchers and lay people appear to assume that liking something "very much" is different from liking something only "a little bit." And indeed, the non-zero between-subjects correlations reported above suggest that people are somewhat calibrated to assess whether their biases are mild or strong compared to one another, albeit to a very limited degree.

To test whether people might be better calibrated in domains where it is both more common to communicate one's preferences and less socially threatening to voice strong preferences than in the domain of social groups, Hahn and Goedderz (in prep. b) asked participants to predict their scores on five IATs towards baked goods. Participants were better calibrated in the domain of baked goods (average r = .39) than in the domain of social groups (average r = .21), despite similar introspective awareness (corrected r's = .55 and .63 for baked goods and social groups, respectively). Additional analyses showed that most participants used only the mildest labels available on the 7-point scales when predicting their IAT scores in the domain of social groups (most participants describe their biases as "slight", Anchors 3 and 5 on the 7-point scales, regardless of strength). In contrast, predictions in the domain of baked goods encompassed the full scale, frequently including "a lot more positive" reactions (Anchors 1 and 7) towards some categories of baked goods than others. In fact, participants predicted the same IAT scores with significantly stronger labels when they measured implicit evaluations of baked goods than when they measured implicit evaluations of social groups. These findings suggest that participants are less well calibrated when the available anchors sound socially undesirable.

We return to the factors that influence calibration below. As it stands, the available data suggest that people tend to be miscalibrated in their biases towards social groups, but that social calibration varies widely as a function of domain, although it might never be perfect.

Discussion

Our analyses suggest that the cognitions reflected on implicit measures should not be called "unconscious" when this term is used as a *trait* that defines them at all time. The available data suggest that it is possible to be aware of them. However, our analyses also suggest that those cognitions might sometimes be *state*-unconscious when people are asked to simulate their biases in the abstract rather than in a concrete encounter; and *pre*conscious until a person pays specific attention to their spontaneous reactions in such concrete encounters.

The description "unconscious attitudes" implies a trait definition that describes the cognitions reflected on implicit measures as unconscious at all times. In contrast, assuming that people are generally "unwilling" to report on those cognitions implies a definition of those cognitions as trait-conscious. Our findings support neither of these definitions. Instead, they suggest that the cognitions reflected on implicit measures may be defined as awareness-independent. They can variably be (state-) conscious, (state-) unconscious, or preconscious at different times. This differentiates implicit from explicit measures, which presumably always reflect conscious cognitions.

Lastly, our review of the evidence suggests that people are often miscalibrated in reporting the cognitions reflected on implicit measures in socially sensitive domains. Hence, they may be socially unaware of where their biases fall compared to others; and they may look unaware of socially undesirable biases when awareness is determined from between-subjects analyses.

State-Unawareness of Everyday Biases?

The findings indicating that people might often be state-unaware of their social-group biases are puzzling in many ways. Participants in the studies reported here can be assumed to have met people with different backgrounds throughout their lives. Why, then, do they appear to need concrete stimuli to notice their biases? And why do they seem to discover new information when they are asked to pay attention to their spontaneous biases? Could they not just remember their reactions in previous situations? We have several ideas for possible answers, all of which require further research. First, in line with Process 1b of our process model, people may simply never pay attention to their reactions in these domains, such that those cognitions remain preconscious in all concrete encounters, and cannot be recalled later. This interpretation is consistent with our findings that, without an attention manipulation, people remain surprised at bias feedback when they complete IATs, even though concrete stimuli are present throughout the entire test (Goedderz & Hahn, in prep.). Second, people may recognize their reactions, but either fail to integrate them into long-term memory, or attribute them to other sources than the background of the targets. These possibilities are consistent with observations that people often notice their different reactions in the different blocks of the IAT, but attribute them to procedural details of the test other than racial bias (e.g., the block order of the test, Monteith et al., 2001). One interpretation of these possibilities is that people ignore their affective reactions to specific social categories because those are at odds with their values, akin to the concepts of repression or suppression (Krickel, 2018; Wilson et al., 2000). A third possibility is that

there might be further moderating factors that increase the strength of the affective reaction (Process 1a) that are present in the research presented here, but not in real-world encounters or during IAT completion. For instance, in our studies, predictions are made by putting two categories in contrast (e.g., "is your reaction more positive towards BLACK or more positive towards WHITE?"). In the real world, people often meet outgroup members individually, but contrast might be necessary to notice one's biased reactions. The current data are compatible with all five of those processes at the same or different times. Future research is needed to explain why people appear to find new information when they discover their social-group biases.

Social Calibration and Social Desirability

Our findings on calibration may be described as a problem of socially desirable responding (SDR, Crowne & Marlowe, 1960). However, we think reducing calibration to SDR would cloud two important components of the calibration process. First, SDR is often used to imply that participants know the "true" answer to a question and hide this answer deliberately and dishonestly (Crowne & Marlowe, 1960). Because participants in Hahn et al's (2014) studies know their biases will be revealed, dishonest reporting would be futile. Hence, participants likely honestly chose to describe their biases with labels they considered appropriate. In other words, social calibration more likely reflects a lack of social awareness than deliberate dishonesty. Although SDR may be interpreted in broad ways to include such an interpretation, social calibration is a more precise description that shows that even fully conscious cognitions can be construed in self-serving ways.

The second reason is that other motivations than SDR likely influence the calibration process. Imagine an egalitarian person who believes that the concept of implicit bias is not appreciated enough in society. This person may show a bias that may be termed "slight" by others, but insist on calling this bias "strong" because of their belief that implicit biases are more meaningful than many people assume. In this example, the person's motivation would lead them to choose the less socially desirable response, but they would nevertheless be miscalibrated if their biases would be labeled mild by others. Social calibration encompasses all processes and motivations that lead people to map an answer they want to report onto a scale, or into any linguistic format. In fact, the low between-subjects correlations in the data by Hahn and Goedderz (in prep. b) and Hahn et al. (2014) show miscalibration in both directions, even if they show a tendency to report low levels of social biases overall. We believe that this emphasis on scale usage makes social calibration both more precise (in that it goes beyond dishonest responding) and more broad (in that it includes motivations beyond SDR) than the concept of SDR. Notwithstanding these thoughts, our findings suggest that people are generally motivated to calibrate their responses in socially desirable ways in the domain of social groups. More research is needed to determine whether our new construct will prove useful in describing reports on automatic cognitions.

What Exactly Are People Introspecting Upon When They Predict Their IAT Results?

We will attempt to answer this question in two ways: methodologically and conceptually. Methodologically, much research has shown that implicit evaluations result from a multitude of processes, many of them unique to the method that is used (Conrey et al., 2005; Klauer et al., 2007; Rothermund & Wentura, 2004). Testing the contribution of method-specific variance to accuracy in IAT score predictions

experimentally, Hahn et al. (2014) found that predictions were if anything non-significantly most accurate when participants had received no explanation on how the IAT works and had no experience with it. These results make it unlikely that participants predicted method-specific variance. On the other hand, participants were more accurate in predicting their IAT scores when they saw precisely the same stimuli as used on the IAT in their predictions, indicating at least some degree of stimulus variance (Hahn & Goedderz, in prep. a). Recall, however, that other concrete stimuli already led to better predictions, such that the stimulus effect explains only part of the results. Together, these findings suggest that people can predict variance in their IAT scores beyond its specific methodology and hence must have some access to the underlying cognitions producing those results. But what exactly are these underlying cognitions?

Rivers and Hahn (2019) used the quadruple process model (Conrey et al., 2005) to determine which of five processes contributing to IAT responses was most strongly related to IAT score predictions in Hahn et al.'s (2014) data. Predictions were best explained by a combination of the difference in associations with one target group as opposed to the other (e.g., positive-White + negative-Black associations) in combination with control processes. These results suggest that the cognitions participants predict may be seen as a summary construct, rather than just one specific component.

Conceptually, our working model assumes that people can gain awareness of automatic cognitions by introspecting upon the phenomenological output these cognitions produce (Hahn & Goedderz, in prep. b; Hofmann & Wilson, 2010). Based on research showing that implicit and explicit evaluations are more strongly related when participants rely on affect for their explicit reports (e.g., Ranganath et al., 2008; for a meta-analysis, see Hofmann, Gawronski et al., 2005), we postulated that the cognitions reflected on implicit evaluations manifest themselves in spontaneous affective reactions (Gawronski & Bodenhausen, 2006; Hahn et al., 2014; Hahn & Gawronski, 2019). And indeed, when Hahn and Goedderz (in prep. a) asked participants to report just their spontaneous "gut responses" without mention of a test, those scores showed relationships with IAT scores that were similar to IAT score predictions.

Does this mean that participants can observe the cognitions reflected in IAT scores directly or does it mean that they infer them from their "gut reactions"? This question is difficult to answer because there is a lack of consensus for how the spontaneously activated cognitions reflected on implicit measures are best conceptualized. While most traditional models refer to the underlying construct of interest as an association (Fazio, 2007; Gawronski & Bodenhausen, 2006; Greenwald & Banaji, 1995) or a combination of activated associations and control (Conrey et al., 2005; Payne, 2005), De Houwer (2014) has argued that the cognitions reflected on implicit measures should be described as automatically activated propositions. This conceptual disagreement makes it difficult to say whether observing a "gut reaction" means observing or inferring the underlying cognition. To us, it sounds equally plausible to propose that the gut reaction *is* the underlying cognition (and one may then discuss how to conceptualize a gut reaction scientifically), as it would be to propose that the underlying cognition is something else that *produces* a gut response. We hope that continuous advancement in theorizing will shed further light on this question.

In sum, our review of the available data suggests that people rely on spontaneous "gut responses" when they predict implicit

evaluations, beyond method-specific variance of the measurement instruments. They further suggest that those reactions can be dissociated into multiple cognitive processes (e.g., the contrast between two associations + impulses to control these associations). More research and theorizing are needed to refine what aspects of the mind specifically people can or cannot directly observe.

Awareness and Attitude Process Models

Many dual-process models declare either that (1) the cognitions reflected on implicit measures are "unconscious" (Greenwald & Banaji, 1995; Nosek & Hansen, 2008), or that (2) people deliberately reject these cognitions for explicit reports (e.g., Fazio, 2007; Gawronski & Bodenhausen, 2006; for a historical analyses of these differences, see Payne & Gawronski, 2010). While the former models assume that implicit measures reflect trait-unconscious cognitions, the latter assume those cognitions are trait-conscious, but often rejected. In contrast to such assumptions, our analyses suggest that the cognitions reflected in IAT scores are generally accessible, but often remain at least preconscious until people pay attention to their spontaneous reactions under specific circumstances. They are neither trait-unconscious nor trait-conscious. Hence, both types of dual-process models would require extensions to accommodate them.

The idea that implicit measures reflect unconscious attitudes per se, or "introspectively unidentified … traces of past experience" (Greenwald & Banaji, 1995, p. 8) is incompatible with our finding that people can predict the patterns of their IAT scores. However, our data are compatible with the notion that people are often unaware of their biases both when they report explicit evaluations and when they complete tests that provide implicit measures (Goedderz & Hahn, in prep.). Despite this compatible aspect, the fact that people are generally able to report on the valence reflected on implicit measures contradicts the notion of two fundamentally distinct systems of evaluation.

Concerning the APE model, our data are compatible with the assertion that an explicit report is formed from the evaluative information that is salient during responding (Gawronski & Bodenhausen, 2006). When participants were encouraged to pay attention to their spontaneous affective reactions, those reactions were significantly more strongly reflected in their explicit reports; and implicit-explicit correlations generally vary across participants (Hahn et al., 2014). However, our data also suggest that people do not always consider the cognitions reflected on implicit measures; and when they are encouraged to consider them, they appear to be finding new information. The APE model may accommodate these findings if the term "reject" is broadened to encompass a lack of consideration rather than a deliberate, conscious rejection. Our data are also compatible with single-system models that assume that an explicit evaluation is formed via an iterative process where different information is considered and re-considered until a person decides on one evaluation to report (e.g., Cunningham et al., 2007).

In sum, our data are hard to reconcile with dual-systems accounts that declare the cognitions on implicit measures to be "unconscious" separate structures (Greenwald & Banaji, 1995). They are more compatible with either interacting dual-process or single-process models. However, many of these models may require specifications concerning how often the cognitions reflected on implicit measures factor into the process of stating an explicit evaluation, instead of assuming they are always considered, but then rejected.

How Useful Are Implicit Measures?

One may wonder whether our findings suggest that implicit measures are superfluous. After all, one could just ask participants about their spontaneous reactions instead of their deliberately endorsed attitudes to get the same information that implicit measures reveal. We believe such an interpretation would be premature for at least three reasons.

First, the within-subjects correlations of .50-.70 are high, but not perfect. Additionally, accuracy varies between participants, and research with other attitude targets might show different values. Although our results contradict the notion that implicit measures capture consciously inaccessible cognitions per se, they could still be capturing variance that is not revealed in self-report measures for all participants and all targets. Second, recall that once "unconscious" is defined as a state rather than a trait of cognitions, the cognitions reflected on implicit measures appear to remain preconscious until a person is asked to pay attention to their spontaneous reactions under specific circumstances (e.g., concrete stimuli shown in contrast). Lastly, even if those circumstances are present, people tend to be miscalibrated in their responses. Reports on spontaneous reactions cannot indicate who is more biased than who when everyone thinks they are less biased than everyone else (Hahn et al., 2014), but this variance is necessary to assess whether spontaneous biases predict behavior (Kurdi et al., 2019). Implicit measures rank people's reactions in ways that they themselves may be unable or unwilling to do.

In sum, we believe our analyses so far contradict the notion that a simple question about spontaneous reactions could replace implicit measures. Implicit measures can reveal information people are temporarily not accessing when asked, or unwilling to report for other reasons, and they rank those cognitions in ways that people may be unable to do themselves.

Practical Implications

We also believe that the data presented here have implications for bias interventions. Assuming that biases are unconscious implies teaching people about them, e.g., with feedback. Assuming that people are unwilling to report on their biases implies making people admit to something they already know. In contrast to both of those ideas, our data on how biases are accessible but often preconscious suggest that people should be encouraged to pay attention to their reactions to concrete stimuli. And indeed, in a series of studies Hahn and Gawronski (2019) showed that encouraging people to pay attention to their biases was more effective in raising acknowledgment of bias than completing IATs. Similarly, our findings that people are often miscalibrated and motivated to construe their biases as weaker than the biases of others (Hahn et al., 2014) suggest that interventions may target not only awareness of bias, but comparative social meaning of biases as well. As described above, however, neither of these conclusions would be possible if one follows a standard definition of implicit measures reflecting "attitudes people are unable or unwilling to report" (e.g., ProjectImplicit.net, 2020). Hence, we believe a more nuanced understanding of awareness and implicit measures is not only valuable for theory, but also paramount for the application of implicit measures to societal questions.

Limitations

The analyses presented here are based on a limited set of studies from our lab run on similar paradigms with evaluative IATs on social groups. We initially chose the IAT because it is the most widely used measurement instrument, because it shows acceptable psychometric properties

(Gawronski & De Houwer, 2014), and later because we wanted to systematically vary one aspect at a time from the original paradigm. Future research with different measures and attitude targets is needed to further validate and refine the ideas presented here. Additionally, our data are limited to implicit evaluations, which we assume manifest themselves in spontaneous affective reactions (Gawronski & Bodenhausen, 2006). This poses the question of whether people would be similarly able to predict less affect-based implicit measures, such as implicit stereotypes. We hope that the distinction between unconscious, awareness-independent, and miscalibrated cognitions we have deduced from our research will be a useful framework to conduct additional research, and develop and refine our ideas.

Conclusion

The cognitions reflected on implicit evaluations are often referred to as "unconscious" attitudes. Our analysis suggests that this is an incorrect characterization when the term "unconscious" is used as a trait that describes those cognitions at all times. However, when "unconscious" is defined as a state in which cognitions can be at specific points in time, then our data are compatible with the notion that the cognitions reflected on implicit measures *can* be unconscious. First, people often do not pay attention to their biased reactions towards social groups, such that they may often unfold preconsciously. Second, it might be more difficult to observe one's reactions under some conditions (e.g., in the abstract) than others (e.g., with concrete stimuli), suggesting that conscious access is sometimes inhibited. Additionally, people may be unwilling or unable to calibrate their biases consistently with social norms, such that even when people observe their biases they draw unrealistic conclusions that their reactions are more acceptable than the reactions of others. Together, these results suggest that people are sometimes not aware of the cognitions reflected on implicit measures, and that there are specific aspects they do not know (i.e., their comparative social standing). Future research is needed to validate whether these ideas will hold up to empirical scrutiny with other measures and targets in other domains.

References

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance. *Journal of Personality and Social Psychology*, *89*(4), 469–487. https://doi.org/10.1037/0022-3514.89.4.469

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. https://doi.org/10.1037/h0047358

Cunningham, W. A., Zelazo, P. D., Packer, D. J., & van Bavel, J. J. (2007). The iterative reprocessing model: a multilevel framework for attitudes and evaluation. *Social Cognition*, *25*, 736–760. https://doi.org/10.1521/soco.2007.25.5.736

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353. https://doi.org/10.1111/spc3.12111

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211. https://doi.org/10.1016/j.tics.2006.03.007

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1–37.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, *32*(2), 188–200. https://doi.org/10.1177/0146167205282152

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25*, 603–637. https://doi.org/10.1521/soco.2007.25.5.603

Gawronski, B. (2019). Six Lessons for a Cogent Science of Implicit Bias and Its Criticism. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *14*(4), 574–595. https://doi.org/10.1177/1745691619826015

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 283–310). Cambridge University Press.

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition*, *15*(3), 485–499. https://doi.org/10.1016/j.concog.2005.11.007

Goedderz, A., & Hahn, A. (in prep.). *Preconscious Attitudes: people are surprised at their IAT scores unless they first observe their own biased reactions* [Manuscript in preparation.]. University of Cologne.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit Social Cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Hahn, A., & Gawronski, B. (2018). Implicit social cognition. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, *4*, 1–33. https://onlinelibrary.wiley.com/doi/full/10.1002/9781119170174.epcn412

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116*(5), 769–794. https://doi.org/10.1037/pspi0000155

Hahn, A., & Goedderz, A. (in prep. a). *Beyond dishonesty and unawareness. Accuracy of IAT score predictions depends more on the concreteness of the question than on knowledge of measurement* [Manuscript in preparation]. University of Cologne.

Hahn, A., & Goedderz, A. (in prep. b). *Unaware or Miscalibrated? Distinguishing Introspective Awareness From Social Calibration in Research on Implicit Attitudes* [Manuscript in preparation]. University of Cologne.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369–1392. https://doi.org/10.1037/a0035028

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

Hofmann, W., Gschwendner, T., & Schmitt, M. (2005). On implicit-explicit consistency: the moderating role of individual differences in awareness and adjustment. *European Journal of Personality*, *19*(1), 25–49. https://doi.org/10.1002/per.537

Hofmann, W., Gschwendner, T., & Schmitt, M. (2009). The road to the unconscious self not taken: Discrepancies between self- and observer-inferences about implicit dispositions from nonverbal behavioural cues. *European Journal of Personality*, *23*(4), 343–366. https://doi.org/10.1002/per.722

Hofmann, W., & Wilson, T. D. (2010). Consciousness, introspection, and the adaptive unconscious. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 197–215). Guildford Press.

Howell, J. L., Collisson, B., Crysel, L., Garrido, C. O., Newell, S. M., Cottrell, C. A., Smith, C. T., & Shepperd, J. A. (2013). Managing the Threat of Impending Implicit

Attitude Feedback. *Social Psychological and Personality Science*, *4*(6), 714–720. https://doi.org/10.1177/1948550613479803

Howell, J. L., & Ratliff, K. A. (2017). Not your average bigot: The better-than-average effect and defensive responding to Implicit Association Test feedback. *British Journal of Social Psychology*, *56*(1), 125–145. https://doi.org/10.1111/bjso.12168

Howell, J. L., Redford, L., Pogge, G., & Ratliff, K. A. (2017). Defensive Responding to IAT Feedback. *Social Cognition*, *35*(5), 520–562. https://doi.org/10.1521/soco.2017.35.5.520

Hütter, M., & Sweldens, S. (2013). Implicit misattribution of evaluative responses: Contingency-unaware evaluative conditioning requires simultaneous stimulus presentations. *Journal of Experimental Psychology. General*, *142*(3), 638–643. https://doi.org/10.1037/a0029989

Hütter, M., Sweldens, S., Stahl, C., Unkelbach, C., & Klauer, K. C. (2012). Dissociating contingency awareness and conditioned attitudes: Evidence of contingency-unaware evaluative conditioning. *Journal of Experimental Psychology. General*, *141*(3), 539–557. https://doi.org/10.1037/a0026477

Kassin, S., Fein, S., & Markus, H. R. (2011). *Social Psychology* (8th ed.). Wadsworth, Cengage Learning.

Kihlstrom, J. F. (2004). Implicit methods in social psychology. In C. Sansone, C. C. Morff, & A. T. Panter (Eds.), *The Sage handbook of methods in social psychology* (195-212). Sage.

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology*, *93*(3), 353–368. https://doi.org/10.1037/0022-3514.93.3.353

Krickel, B. (2018). Are the states underlying implicit biases unconscious? – A Neo-Freudian answer. *Philosophical Psychology*, *31*(7), 1007–1026. https://doi.org/10.1080/09515089.2018.1470323

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(13), 5862–5871. https://doi.org/10.1073/pnas.1820240116

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing Implicit Prejudice. *Social and Personality Psychology Compass*, *7*(5), 315–330. https://doi.org/10.1111/spc3.12023

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. D. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). Academic Press.

McConnell, A. R., Dunn, E. W., Austin, S. N., & Rawn, C. D. (2011). Blind spots in the search for happiness: Implicit attitudes and nonverbal leakage predict affective forecasting errors. *Journal of Experimental Social Psychology*, *47*(3), 628–634. https://doi.org/10.1016/j.jesp.2010.12.018

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, *19*(4), 395–417. https://doi.org/10.1521/soco.19.4.395.20759

Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes & Intergroup Relations*, *8*(1), 39–52. https://doi.org/10.1177/1368430205048615

Norman, E. (2010). "The Unconscious" in Current Psychology. *European Psychologist*, *15*(3), 193–201. https://doi.org/10.1027/1016-9040/a000017

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565–584. https://doi.org/10.1037/0096-3445.134.4.565

Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science*, *16*(2), 65–69. https://doi.org/10.1111/j.1467-8721.2007.00477.x

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115. https://doi.org/10.1037//1089-2699.6.1.101

Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition & Emotion*, *22*(4), 553–594. https://doi.org/10.1080/02699930701438186

Payne, B. K. (2005). Conceptualizing control in social cognition: How executive functioning modulates the expression of automatic stereotyping. *Journal of Personality and Social Psychology*, *89*(4), 488–503. https://doi.org/10.1037/0022-3514.89.4.488

Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 1–15). Guildford Press.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice. *Psychological Inquiry*, *28*(4), 233–248. https://doi.org/10.1080/1047840X.2017.1335568

Powell, J. A. (2016, September 27). *Implicit bias in the presidential debate [Blog post]*. UC Berkeley. Berkeley Blog. https://blogs.berkeley.edu/2016/09/27/implicit-bias-in-the-presidential-debate/

ProjectImplicit.net. (2020). https://implicit.harvard.edu

Quillian, L. (2008). Does unconscious racism exist? *Social Psychology Quarterly*, *71*(1), 6–11. https://doi.org/10.1177/019027250807100103

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, *44*(2), 386–396. https://doi.org/10.1016/j.jesp.2006.12.008

Rivers, A. M., & Hahn, A. (2019). What Cognitive Mechanisms Do People Reflect on When They Predict IAT Scores? *Personality & Social Psychology Bulletin*, *45*(6), 878–892. https://doi.org/10.1177/0146167218799307

Rothermund, K., & Wentura, D. (2004). Underlying processes in the implicit association test: Dissociating salience from associations. *Journal of Experimental Psychology. General*, *133*(2), 139–165. https://doi.org/10.1037/0096-3445.133.2.139

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*(6), 995–1008. https://doi.org/10.1037/0022-3514.91.6.995

Schlachter, S., & Rolf, S. (2017). Using the IAT: how do individuals respond to their results? *International Journal of Social Research Methodology*, *20*(1), 77–92. https://doi.org/10.1080/13645579.2015.1117799

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science (New York, N.Y.)*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–126. https://doi.org/10.1037//0033-295X.107.1.101

Tables

Table 1
Within-subjects Correlations of IAT Scores With IAT Score Predictions and Traditional Explicit Evaluations (Thermometer Ratings) in Different National Contexts

	IAT Score Predictions				Explicit Evaluations	
	Average	Median	Corrected Average[1]	Average Controlling for Explicit	Average	Average Controlling for Pred.
United States (Hahn et al., 2014), All Studies, *N* = 430	.55	.68	.67	.56	.20	-.03
Germany (Hahn & Goedderz, in prep. b), combined sample from other papers, *N* = 359	.52	.65	.63	.48	.36	.09

Note. The predicted IATs included a Black-White IAT, a Latino-White IAT, an Asian-White IAT, a celebrity-regular IAT, and an adult-child IAT. Explicit evaluations were thermometer ratings of Whites, Black people, Latinos, Asians, celebrities, regular people (non-celebrity), children, and adults, ranging from 0 – "cold and unfavorable feelings" to 100 – "very warm and favorable feelings", collapsed into differences scores matching the IATs. Predictions were made on 7-point scales and included the pictures and group labels used on the IATs. The specific questions and explanations in the predictions were varied in Hahn et al. (2014), but these variations produced no meaningful differences. All correlations are significant at $p < .001$, except explicit evaluations predicting IAT scores controlling for predictions in the US samples (upper right-most correlation).

[1] To account for the skew in the distribution of correlations, corrected averages are calculated by Fisher *z*-transforming all correlations, computing their average, and back-transforming this average into a correlation. The corrected averages for Hahn et al. (2014) were calculated for the present paper and is not reported in the original paper. The distribution of correlations between explicit measures and IAT scores were not significantly skewed such that no transformation is reported.